Local versus Global Texture Analysis for Lung Nodule Image Retrieval

Ryan Datteri[a], Daniela Raicu[b], Jacob Furst[b]

[a]Gonzaga University, Spokane WA, USA, 99258

[b]Intelligent Multimedia Processing Laboratory

School of Computer Science, Telecommunications and Information Systems,

DePaul University, Chicago IL, USA, 60604

ABSTRACT

Intensity overlap often occurs in medical images, making it difficult to identify different anatomical structures using intensity alone. Research studies have shown that texture is an important component in quantifying the visual appearance of anatomical structures, and is therefore valuable in the analysis, interpretation, and retrieval of lung nodules.

The goal of our research study is to present a comparison between the different texture models: Gabor filters, Markov Random Field (MRF), and global & local co-occurrence. For comparison purposes we utilized Manhattan, Euclidean, and Chebyshev distances for one-dimensional feature vectors (global co-occurrence) while for two-dimensional feature comparison (local co-occurrence, Gabor filters, and MRF) we utilized the similarity measures Chi-Square and Jeffrey-Divergence. Local co-occurrence contains many different variable aspects in its design that can considerably change the success of its results. A thorough examination of local co-occurrence's variables is discussed.

All of the discussed texture models are presented in the context of our previous Content-Based Image Retrieval (CBIR) System [1]. BRISC utilizes the Lung Image Database Consortium (LIDC) database. We have found that Gabor and MRF texture descriptors produce the best retrieval results regardless of the nodule size, number of retrieved items or similarity metric with an average precision of 88%. Global co-occurrence performed the worse at 44% precision yet when co-occurrence was performed locally (local co-occurrence) the precision results improved to 64%. A combination of all the features worked the best with 91% precision.

Keywords: CBIRS, lung nodules, CT scans, co-occurrence, Gabor filters, Markov Random Fields.

1. INTRODUCTION

There were an estimated 160,390 deaths in the United States due to lung cancer in 2007 [2] and lung cancer accounts for around 29% of all cancer deaths [3]. In this paper we present a CBIR system, BRISC, for pulmonary nodule lookup and similarity retrieval. BRISC will be utilized by radiologists to increase their ability to diagnose pulmonary nodules by retrieving similar cases and allowing the radiologist to directly compare the questionable nodule to other cases.

The hypothesis is that the uncertainty of the radiologist in identifying suspicious lesions can be reduced by providing a visual comparison of a given lesion to a collection of similar lesions of known pathology. To test this hypothesis, we propose to develop a CBIR system whose similarity results match human perception. The human observer (radiologist) manually (or semi-automatically or automatically) segments a lesion from a clinical case. The system computes a set of quantitative descriptors for that lesion and compares those descriptors to the descriptors of known lesions. The underlying assertion is that if a known malignant lesion has certain computable features then unknown lesions with similar computable features would be malignant. Simply put, the expected outcome is a system that provides a way of "looking-up" an image in a collection of images such that similar images are retrieved.

BRISC was previously implemented by Lam et al [1] and the code is also available as open source [4]. Lam at al. showed that the global co-occurrence texture model performs worse (retrieval precision of 29%) than the Gabor filters and MRF texture models (retrieval precision of 88%). In this paper, we implement the co-occurrence texture model at the local level (within a small neighborhood for each pixel of a lung nodule image instead of the entire nodule image); furthermore, we investigate the effect of using all three texture models with respect to their similarity retrieval power.

Given that each one of the texture models captures different properties of the texture, in this paper we show that the combination of the three texture models produces better results than the individual texture models.

There are several CBIR projects in the medical field currently underway. Kinoshita et al. [5] discussed a CBIR system for mammograms. They utilized a large amount of visual features including shape, texture, and granulometric. Kinoshita et al. combined many different features using principal component analysis to improve the system. They report a precision of around 85%. However, based upon the ground truth that they used (BI-RADS categorization) a random selection of images would return a precision of 62.5%. Wei. et al. [6] discussed image analysis and image retrieval using gray level co-occurrence matrices for the mammography domain. They used a distance of 5 on the co-occurrence matrices which resulted in a maximum precision of 51% and recall of 19% (average = 49% and 18% respectively). They grouped each region of interest into 6 categories and if the returned image was placed in the right category it was relevant.

Furthermore, Muramatsu et al. [7] presented strong evidence that CBIR systems could help to improve the accuracy of identifying benign or malignant clustered micro-calcifications on mammograms and that breast radiologists are able to provide a reasonable ground truth. Muramatsu [8] discusses the development of an Artificial Neural Network that shows promising ability to retrieve images similar to those of an unknown lesion. Tourassi et al. [9] evaluated similarity measures utilized in a scheme for content-based retrieval and detection of masses in screening mammograms. They found that the measures interestingly fell into two categories: "one category is better suited to the retrieval of semantically similar cases while the second is more effective with knowledge-based decisions regarding the presence of a true mass in the query location". Zheng et al. [10] has also done interesting work developing an automated interactive computer-aided diagnosis scheme that performs just as well as the subjective rating method.

Computed Tomography (CT) scanning has been found to increase the detection rate of pulmonary nodules [11]. Much work has been done to develop computer assisted diagnosis and detection (CAD) systems for pulmonary nodules in CT. For a detailed description of CAD systems, we suggest the review by Muller et al [12].

One of the largest CBIR projects currently underway using lung CT images is the ASSERT project [13], which is being developed at Purdue University and was first published in 1999. It proposed a "physician-in-the-loop" system where radiologists highlight a region and the system would return similar images. The system used a variety of image features including co-occurrence statistics, shape descriptors, Fourier transforms and global gray level statistics. The system also utilized physician-provided ratings of features such as homogeneity, calcification and artery size. The best precision reported by the system was 76.3%.

There are many difficulties involved with content-based retrieval of medical images, including the difficulty of automatic segmentation, the large variability of feature selection, and the lack of standardized toolkits and evaluation methods [14][15][16]. There have been efforts recently to solve these problems including the Lung Image Database Consortium (LIDC) collection which was specifically developed to support evaluation and comparison of chest CAD systems [17]. It can be used similarly to develop, evaluate, and compare CBIR systems.

2. METHODOLOGY

2.1 LIDC Data

The data in our study was obtained from the Lung Image Database Consortium (LIDC) database [17]. The database contains 149 unique pulmonary nodules that have been segmented and annotated by up to four different radiologists amounting to a total of 2020 images. These images were taken from a total of 90 Computer Tomography studies of the chest, each containing between 100 and 400 Digital Imaging and Communication (DICOM) images. Four radiologists marked the contour of the nodules and assigned nine semantic terms/characteristics to each nodule: calcification, internal structure, lobulation, malignancy (as interpreted by the radiologists based on imaging findings), sphericity, spiculation, subtlety, texture, and margin. Calcification and internal structural are nominal while the other seven annotations are ordinal. Calcification contains six different categories, internal structure contains four different categories, and the other seven annotations each are rated on a scale from one to five.
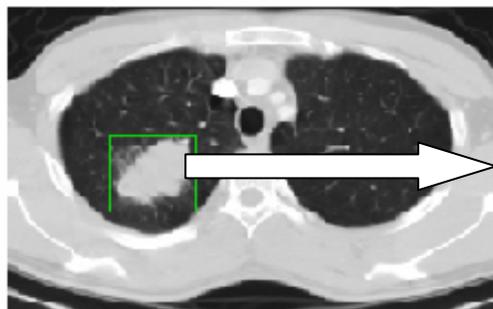
[calcification = 6, internal structure = 1, lobulation = 3, malignancy = 5, sphericity = 3, speculation = 3, subtlety = 5, texture = 5, and margin = 3]

Fig. 1. Example CT slice with nodule

2.2 Texture Models

We extract low-level image features that encode the *texture* of the lung nodules while satisfying the main requirements for feature extraction: a) *completeness/expressiveness* (features should be a rich enough representation of the image contents to reproduce the essential information); b) *compactness* (the storage of the features should be compact to allow efficient access) and c) *tractability* (the distance between features should be efficient to compute).

The three texture models satisfying these properties and used for this research are: local and global co-occurrence matrices [18], Gabor filters [21], and MRF [18][22]. The co-occurrence texture models generated 11 texture descriptors which represented the statistical properties of the nodules' texture. Separate co-occurrence matrices were calculated for each direction (0, 45, 90, and 135 degrees) and displacement (1, 2, 3, and 4 pixels). In global co-occurrence the texture descriptors were extracted per nodule image while in local co-occurrence the texture descriptors were extracted for each relevant pixel in the nodule image. The intensities of the nodule image were binned for global and local co-occurrence to allow statistical relevance to appear in the co-occurrence matrices; otherwise the information gained is usually noise. For pixel level feature extraction local co-occurrence extracts a set of pixels that surrounds each pixel and performs co-occurrence on that subset of the original image. The size of that subset is determined by the variable 'window size'. Also, to compute similarity, local co-occurrence is placed into a histogram. The variables for window size, number of bins used for the histogram, and the number of bins for the intensities were varied in an attempt to find the parameters that achieved the best results in local co-occurrence.

In contrast to the statistical based co-occurrence methods, Gabor filtering is a transform-based method of extracting texture information in the form of a response image. A Gabor filter is a sinusoid function modulated by a Gaussian and produces 12 filter images tuned to four orientations (0, $\pi/2$, $\pi/4$, $3\pi/4$) and three frequencies (.3, .4, and .5) encoding the texture properties in the frequency space [1]. Markov Random Fields capture the local contextual information of an image. The value utilized for each pixel in MRF is dependant on its neighbors. The MRF model produced five images corresponding to four orientations (0, 45, 90, and 135 degrees) between pairs of neighboring pixels plus variance [1].

2.3 Similarity Measures

There are many similarity measures proposed for general CBIR systems and the choice of a similarity measure is dependent on both the feature space representation and its ability to capture the visual human perception of similarity. We investigate similarity measures from three categories of similarity measures: 1) *Heuristic distance metrics* (Minkowski distance), 2) *Non-parametric test statistics* (Chi-square statistics), and 3) *Information Theory Divergences* (Jeffrey-Divergence).

Global co-occurrence results in a one dimensional feature vector for each image, therefore Euclidean, Manhattan, and Chebyshev were used to measure the texture-based similarity of the nodules. Local co-occurrence, MRF, and Gabor features are local, so they result in a two dimensional feature response for each image. Thus the Chi-Square and Jeffrey-Divergence measures were used to measure the texture-based similarity between these models. More information about these similarity measures can be found in the papers by Lam et al [1] and Puzicha et al. [19], [20].

2.4 Retrieval Performance Evaluation

We evaluate the retrieval system using precision as the performance metric. The precision is calculated for all images from the database; the overall precision of the system is then calculated as the average of all precision values obtained when each image becomes the query image. The general formula for calculating precision is:

$$\mathrm{Pr}ecision = \frac{\#\_of\_relevant\_images}{\#\_of\_retrieved\_images}$$

We calculate the precision considering as a "relevant image" an image belonging to the same nodule but appearing in another slice or even in the same slice but outlined by another radiologist (see Figure 2).
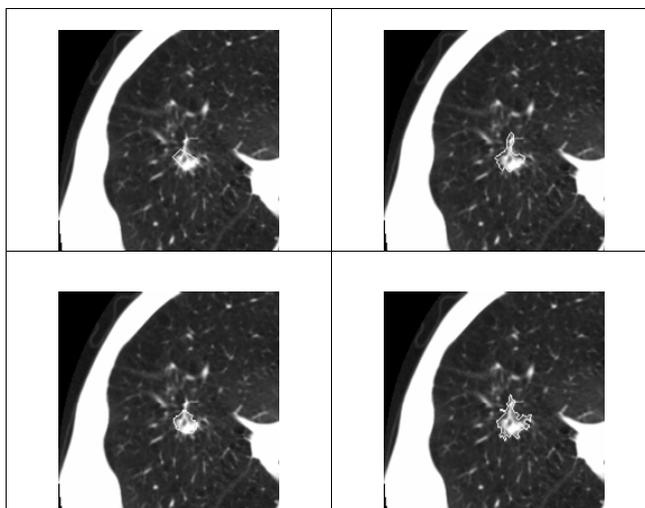

Fig. 2. Four distinct outlines for a nodule delineated by four radiologists

3. RESULTS

3.1 Local co-occurrence parameters

For local co-occurrence we tested utilizing the window sizes (5x5, 7x7, 9x9, 11x11, and 13x13), the number of bins for intensities (64, 128, 256, and 512), and the number of bins used for the histogram representation of the local texture descriptors (8, 16, 32, 64, 80, and 96). The best combination of the three variables was 5x5 window size, 64 bins for intensities, and 96 bins used for the histogram. We did not use window size 3x3 because previous literature indicates that other methods would be more appropriate [23]. These are the values that we utilized for local co-occurrence during the testing for the following results.

3.2: Local co-occurrence vs. global co-occurrence

Figure 3 shows the results for local co-occurrence when using the optimal values (i.e. 5x5 window size, 64 bins for intensities, and 96 bins for the histogram) that we found compared with global co-occurrence. To calculate similarities we utilized Jeffrey-Divergence for local co-occurrence and Euclidean for global co-occurrence. Local co-occurrence performs nearly 20% better for 1 item retrieved in response to the query, 17% for 2 items, 13% for 3 items, 9% for 5 items and 6% better when 10 items are retrieved. Co-occurrence therefore performs much better when it gathers texture information at the pixel level.
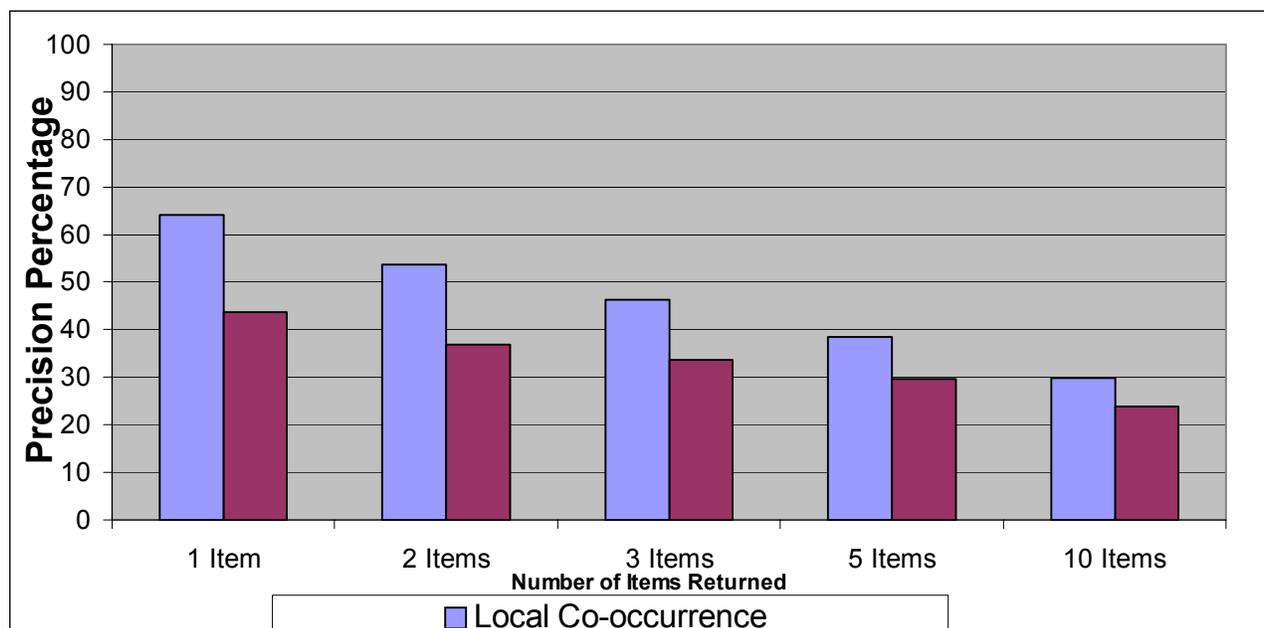
Fig 3: local co-occurrence vs. global co-occurrence

3.2 Texture Comparison

Figure 4 shows the best result for one item returned for global co-occurrence, local co-occurrence, Gabor filters, and MRF. Gabor filters and MRF both utilized the similarity measure Chi-Square while local and global co-occurrence utilized the similarity measures previously mentioned. Gabor and Markov perform the best with one item retrieved with 88% precision. Local co-occurrence performs relatively well with 64% precision, while global co-occurrence performed the worst with 44% precision. The global co-occurrence precision was improved by 15% over the precision received by Lam [1] because we binned the intensities of the images before performing global co-occurrence, reducing noise and making the method more effective.
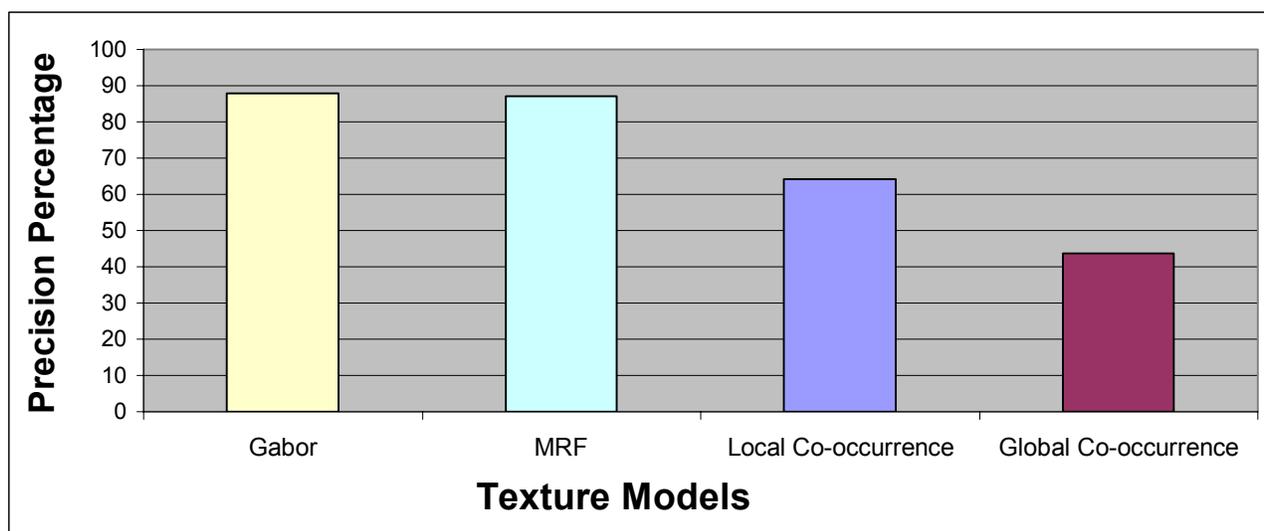

Fig. 4. Texture Comparison: 1 Item Retrieved

3.3 Combination of Features

Fig. 5 shows the relationships between the different texture models including the combination of features (which consists of 205 separate features: Gabor filters use 24 separate features, MRF uses a total of 5 features while local co-occurrence uses 176 features), varying the number of items returned. As one can see, the combination of texture models performs the best (91%) followed closely by Markov and Gabor (88%) with local co-occurrence performing behind them (64%). Global co-occurrence performed much worse then the rest of the texture models (44%). As more items are retrieved the distance between the combination of texture models and the other texture models increases.
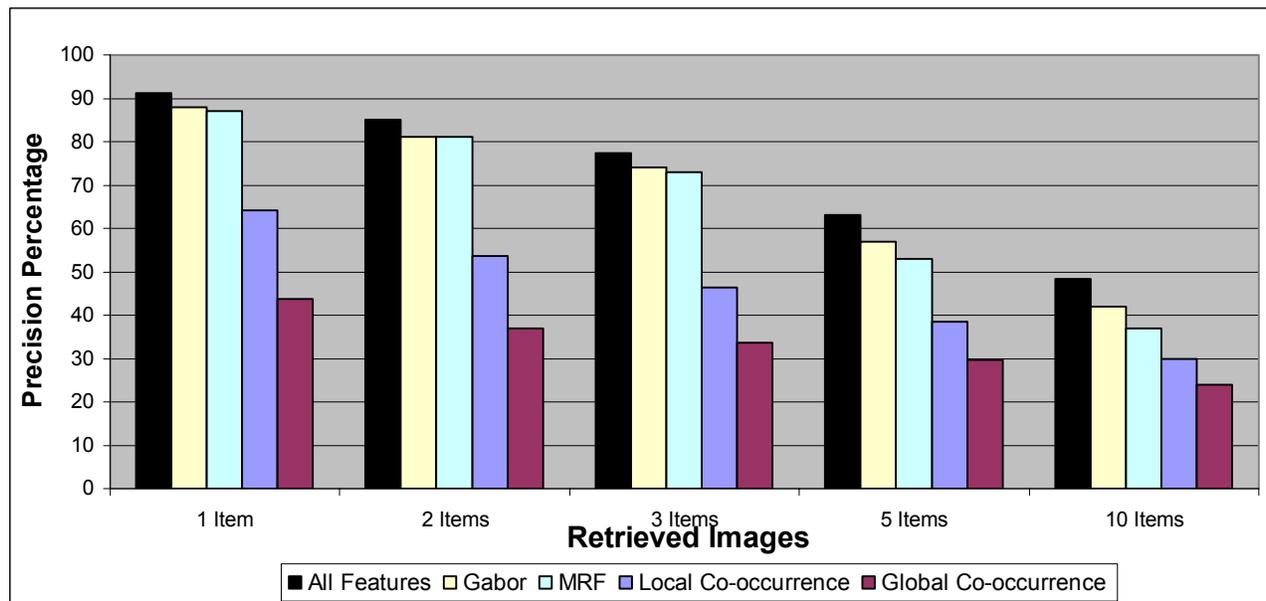

Fig. 5. Images Retrieved Comparison

3.4 Physician Agreement

As discussed previously, four different radiologists annotated each nodule image using a categorical system based upon semantic descriptions of the nodule. We then limited the database of images from which our calculations are made using only the images in which radiologists agreed on the "texture" annotation and ran precision calculations (see Figure 6). When just two radiologists agreed the average precision for MRF and Gabor increased from 88% to 96%, local co-occurrence's precision increased from 64% to 67%, and global co-occurrence's precision had a minimal increase of .2%. Gabor and MRF had nearly 100% precision when three or more radiologists agreed and the global and local co-occurrence method's precision increased by 10% to 54% and 73% respectively when four radiologists agreed.
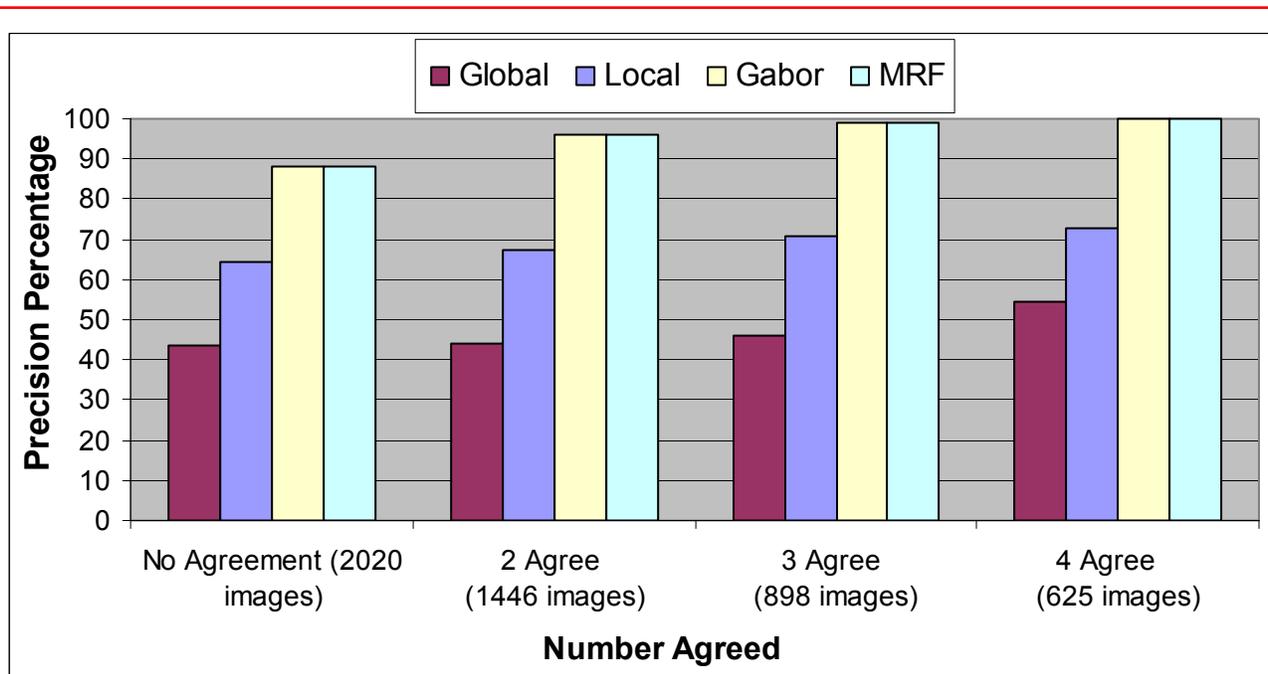
Fig. 6. Physician Agreement

3.5 Image Sizes

Figure 7 shows the results when we ran precision calculations for different nodule image sizes. The size of the nodule image is determined by the number of pixels included within the contours specified by a radiologist. The nodule database was divided into four equal groups based on the size of the nodule images and precision calculations were run with one item retrieved. Methods generally performed better on larger images, except for a decrease in precision in the third group (235-625 total pixels) for MRF and Gabor. Figure 7 shows that local co-occurrence performs as well as Gabor and MRF for larger images but performs noticeably poorer for smaller images (25 – 234 total pixels). Therefore local co-occurrence performs much better when there exists enough data from which it can extract information.
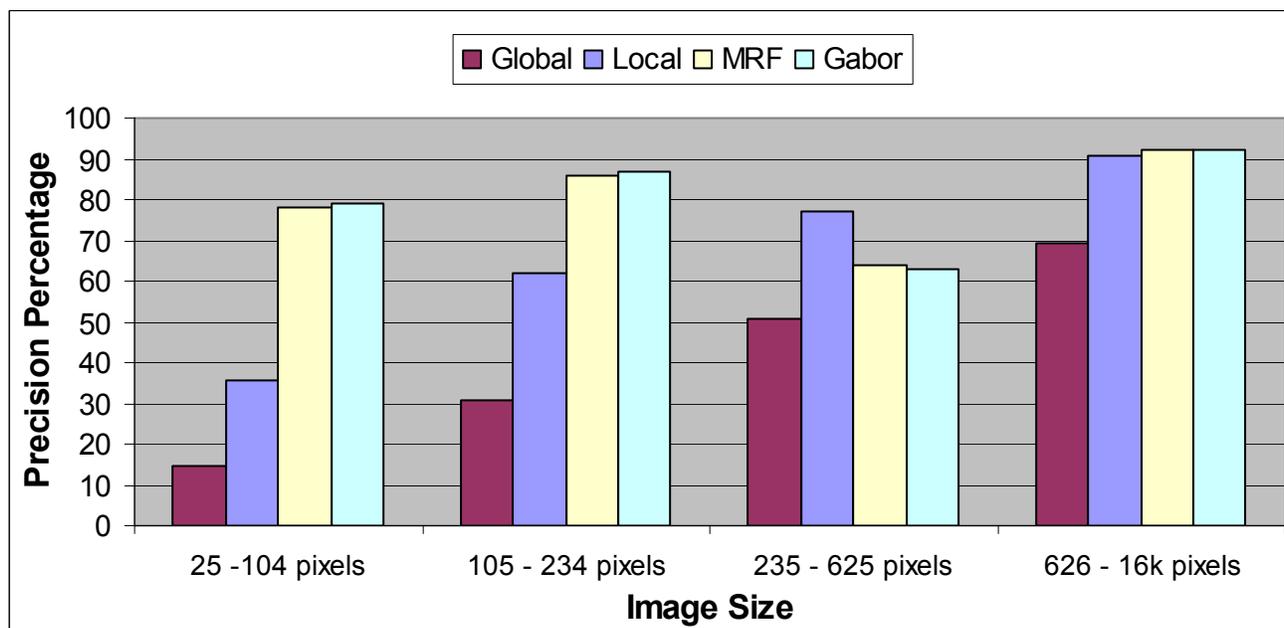

Fig. 7. Image Sizes

4. SUMMARY

In this paper we have shown a comprehensive view and comparison of multiple texture models as well as techniques for improving the texture models. Pixel-level feature extraction performs much better then global extraction, even when dealing with small areas. Gabor and MRF perform the best with local co-occurrence performing comparably while global co-occurrence performs poorly. The combination of the local co-occurrence, Gabor filters and MRF texture models produces the best results consistently. Global co-occurrence was not included in the combination since its features are extracted at the global level while the other three methods extract their features at the local level. In general greater physician agreement led to a greater precision, showing that there is a correlation between radiologists' perception and the texture models. Precision also increases when the size of the image increases, from which we can draw the conclusion that larger lung nodules result in a larger amount of beneficial textural information. In terms of similarity, there is little difference between the Euclidean, Manhattan, and Chebychev although Euclidean repeatedly performs slightly better followed closely by Manhattan, then by Chebychev. Jeffrey-Divergence and Chi-Square perform exactly the same for local co-occurrence yet Chi-Square performs slightly better for Gabor and MRF consistently. The best individual texture models are Gabor filters and Markov Random Field using the similarity measure Chi-Square while the combination of all texture models performed the best.

REFERENCES

[1] M. Lam, T. Disney, M. Pham, D. Raicu, J. Furst, R. Susomboon, "Content-Based Image Retrieval for Pulmonary Computed Tomography Nodule Images", SPIE Medical Imaging Conference, Vol. 6516, 65160N, March, 2007.

[2] *Cancer Facts and Figures*, American Cancer Society, 2006.

[3] "What are the Key Statistics about Lung Cancer?" American Cancer Society, 2007. http://www.cancer.org/docroot/CRI/content/CRI_2_4_1x_What_Are_the_Key_Statistics_About_Lung_Cancer_15.asp?sitearea=.

[4] Lam M, Disney *T*, Raicu DS, Furst JD, Channin DS. "BRISC - An Open Source Pulmonary Nodule Image Retrieval Framework". Journal of Digital Imaging, 2007, doi 10.1007/s10278-007-9059-y.

[5] S. K. Kinoshita, P.M. de Azevedo-Marques, R.R. Pereira Júnior, J.A.H. Rodrigues, and R.M. Rangayyan, "Content-based retrieval of mammograms using visual features related to breast density patterns". Vol 20; number 2, pages 172-190, February 2007.

[6] C. Wei, C. LI, R. Wilson. "A General Framework for Content-Based Medical Image Retrieval with its application to Mammograms", Proc. SPIE Medical Imaging 2005, April 2005; 134-143.

[7] Muramatsu C. Li Q. Schmidt R. Suzuki K. Shiraishi J. Newstead G. Doi K. "Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results". *Medical Physics. 33(9):3460-8, 2006 Sep.*

[8] Muramatsu C. Li Q. Suzuki K. Schmidt RA. Shiraishi J. Newstead GM. Doi K. "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results". Medical Physics. 32(7):2295-304, 2005 Jul.

[9] Tourassi GD. Harrawood B. Singh S. Lo JY. Floyd CE. "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms". Medical Physics. 34(1):140-50, 2007 Jan.

[10] Zheng B. Mello-Thoms C. Wang XH. Abrams GS. Sumkin JH. Chough DM. Ganott MA. Lu A. Gur D. "Interactive computer-aided diagnosis of breast masses: computerized selection of visually similar image sets from a reference library". Academic Radiology, Volume 14, Issue 8, Pages 917-927, 2007 Aug.

[11] Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP: "Early lung cancer action project: overall design and findings from baseline screening". The Lancet 354:99-105, 1999, July.

[12] Muller H, Michoux N, Bandon D, Geissbuhler A: "A review of content-based image retrieval systems in medical applications – clinical benefits and future directions". International Journal of Medical Informatics 73(1):1-23, 2004, February.

[13] C.-R. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," Computer Vision and Image Understanding 75, pp. 111=132, July/August 1999.

[14] Smeulders AW, Worring M, Santini S, Gupta A, Jain R: "Content-based image retrieval at the end of the early years". IEEE Trans Pattern Anal Mach Intelligence 22(12):1349-1380, 2000, (December).

[15] Antani S, Long LR, Thoma GR: "Content-based image retrieval for large biomedical image archives". In Proceedings of 11th World Congress on Medical Informatics (MEDINFO) 2004 (September).

[16] Muller H, Michoux N, Bandon D, Geissbuhler A: "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions". International Journal of Medical Informatics 73(1):1-23, 2004 (February).

[17] LIDC Lung Nodule Image Database. National Cancer Imaging Archive (https://imaging.nci.nih.gov/ncia/). Accessed July 2007.

[18] A. Materka and M.Strzelecki, "Texture analysis methods – a review," tech. rep., Technical University of Lodz, Institute of Electronics, 1998. COST B11 report.

[19] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), 1997.

[20] J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in ICCV (2), pp. 1165-1172, 1999.

[21] T. Andrysiak and M. Choras, "Image retrieval based on hierarchical gabor filters," International Journal Applied Computer Science 15 (4), pp. 471-480, 2005.

[22] C. Chen, L. Pau, and P. W. (eds.), The Handbook of Pattern Recognition and Computer Vision (2nd Edition), World Scientific Publishing Company, 1998.

[23] Malik, J., Belongie, S., Shie, J., and Leung, T. 1999. "Textons, contours and regions: Cue integration in image segmentation". In Proc. IEEE Intl. Conf. Computer Vision, Vol. 2, Corfu, Greece, pp. 918-925.